BMJ Open

Novel risk stratification algorithm for estimating the risk of death in patients with relapsed multiple myeloma: external validation in a retrospective chart review

Roman Hájek ,[1] Sebastian Gonzalez-McQuire,[2] Zsolt Szabo,[2] Michel Delforge,[3] Lucy DeCosta,[4] Marc S Raab,[5] Walter Bouwmeester,[6] Marco Campioni,[2] Andrew Briggs[7]

To cite: Hájek R, Gonzalez-McQuire S, Szabo Z, *et al*. Novel risk stratification algorithm for estimating the risk of death in patients with relapsed multiple myeloma: external validation in a retrospective chart review. *BMJ Open* 2020;**10**:e034209. doi:10.1136/bmjopen-2019-034209

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-034209).

Received 11 September 2019
Revised 28 February 2020
Accepted 28 April 2020

For numbered affiliations see end of article.

Correspondence to
Professor Roman Hájek;
roman.hajek@fno.cz

ABSTRACT

Objectives and design A novel risk stratification algorithm estimating risk of death in patients with relapsed multiple myeloma starting second-line treatment was recently developed using multivariable Cox regression of data from a Czech registry. It uses 16 parameters routinely collected in medical practice to stratify patients into four distinct risk groups in terms of survival expectation. To provide insight into generalisability of the risk stratification algorithm, the study aimed to validate the risk stratification algorithm using real-world data from specifically designed retrospective chart audits from three European countries.

Participants and setting Physicians collected data from 998 patients (France, 386; Germany, 344; UK, 268) and applied the risk stratification algorithm.

Methods The performance of the Cox regression model for predicting risk of death was assessed by Nagelkerke's R^2, goodness of fit and the C-index. The risk stratification algorithm's ability to discriminate overall survival across four risk groups was evaluated using Kaplan-Meier curves and HRs.

Results Consistent with the Czech registry, the stratification performance of the risk stratification algorithm demonstrated clear differentiation in risk of death between the four groups. As risk groups increased, risk of death doubled. The C-index was 0.715 (95% CI 0.690 to 0.734).

Conclusions Validation of the novel risk stratification algorithm in an independent 'real-world' dataset demonstrated that it stratifies patients in four subgroups according to survival expectation.

Strengths and limitations of this study

- The risk stratification algorithm was validated in a real-world dataset containing information on patients with symptomatic multiple myeloma from France, Germany and the UK.
- Using real-world data from the Czech Registry of Monoclonal Gammopathies, HRs for independent predictors of overall survival were derived from a Cox model and individual patient scores were calculated for total risk.
- The performance of the Cox regression model for predicting risk of death was assessed by Nagelkerke's R^2, goodness of fit and the C-index.
- A comparison of the HRs across validation datasets was used to indicate the extent to which scores could be reliably interpreted in different countries/settings.
- The current analysis is limited in that it used a mix of Cox model performance (R^2, C-index) and simply computed HRs to compare risk groups; currently, there are no established measures for assessing risk stratification algorithm performance.

INTRODUCTION

The management of multiple myeloma (MM) can be challenging owing to the heterogeneous nature of patient's characteristics, the disease course and the array of treatment regimens that patients may receive.[1 2] Although many patients respond well to current first-line (1L) treatments, these are not curative and most patients will relapse or become refractory.[3 4] A report from a European chart review has found that while a high proportion of patients with MM received at least one line of active treatment (95%), this decreased significantly with further lines of treatment; 61% received a second line and 15% received a fourth or further line.[5] The type of treatment and treatment sequence also varied considerably at each line.[5]

Defining the prognosis of patients with MM is increasingly challenging. Physicians may consider a range of patient-related and disease-related factors when trying to assess MM prognosis. These can include age, time to progression, best response achieved, cytogenetics and Eastern Cooperative Oncology

Group (ECOG) performance status, as well as comorbidities, previous treatment history (including efficacy and tolerability) and the type of relapse.[2 6] In addition to these factors, validated prognostic and predictive tools are needed in MM to standardise risk stratification of patients and ultimately improve risk assessment.

The International Staging System (ISS) and the revised ISS (R-ISS) have been developed to indicate prognosis in MM and are based on the strongest known predictors at diagnosis.[7–9] The ISS uses a three-stage classification at diagnosis to predict survival based on serum albumin and serum β-2 microglobulin (Sβ2M) levels[7 8] and the R-ISS includes cytogenetic abnormalities (CA) and lactate dehydrogenase (LDH) levels in addition to Sβ2M and serum albumin levels to refine the definition of the disease stage.[9 10] Neither of these tools, however, fully reflect the typical clinical approach to assessing patient prognosis at first relapse[11] nor do they address the experience of the patient during newly diagnosed MM and characteristics of the patient as they relapse. The ISS and R-ISS are therefore less relevant for holistic risk assessment in this setting. The ISS and R-ISS were designed to predict survival based on parameters measured at diagnosis.[7 9] Although the prognostic value of the R-ISS has been demonstrated in both newly diagnosed patients and those with relapsed or refractory MM,[12 13] it does not take account of important indicators available to physicians at first relapse such as the efficacy and safety of 1L treatment. In addition, patient characteristics at initiation of second-line (2L) treatment differ significantly from those at diagnosis (eg, many patients die during frontline therapy). Therefore, there is a need for specifically designed risk assessment tools.[14]

Given the need for specifically designed tools to aid medical decision-making in MM at first relapse, a novel risk stratification algorithm has been developed using real-world data from the Czech Registry of Monoclonal Gammopathies (RMG) for patients with relapsed MM initiating 2L treatment.[15 16] The risk stratification algorithm uses 16 predictors to stratify patients into four risk groups with profoundly different survival expectations. It is the first tool that was designed to include both frailty assessment and disease aggressiveness in a single algorithm to reflect holistic patient-specific risk assessment.

Two manuscripts have recently been published to describe the development of the risk stratification algorithm. Hájek *et al* provided an overview of the algorithm development and addressed how the results can be interpreted from a clinical decision-making perspective.[16] Recently, a manuscript describing in detail the methodology used to develop the risk stratification algorithm has been published.[15] The methodology manuscript explains the processes involved in constructing risk stratification and prognostic tools in oncology. To provide insight into the generalisability of this new risk stratification algorithm and its potential for use in clinical practice, external validation was required to evaluate its performance in independent datasets. A bespoke retrospective chart review cohort study was conducted to collate data from patients in France, Germany and the UK for this purpose. Individual data from these three countries and the pooled validation population were compared with the development cohort and differences in disease characteristics and treatment patterns were examined. In this manuscript, we report the validation of the risk stratification algorithm in detail using real-world data sources from France, Germany and the UK, with an assessment of its performance to predict the risk of death in patients with relapsed MM.

METHODS

Description of the risk stratification algorithm

A Cox model was developed using a conceptual model in MM which combined a systematic literature review and physician judgement (using a Delphi process) to select candidate predictors, followed by a backward selection process using Akaike's information criterion to identify the independent predictors of overall survival (OS).[17] HRs for OS of each predictor were derived from the Cox regression analysis. Risk scores were then calculated by multiplying the HRs for each predictor. The patient-specific score was used as a single variable to stratify patients into four risk groups; these were defined using the K-adaptive partitioning algorithm (total risk score ≤3.0, group 1; >3.0 to ≤7.0, group 2; >7.0 to 15.4, group 3; >15.4, group 4).

The risk stratification algorithm incorporated the following as predictors of OS that are available in routine clinical practice: age, albumin level, bone marrow plasma cell count, thrombocyte count, Sβ2M level, Sβ2M level at diagnosis, LDH level, LDH level at diagnosis, calcium level, time to next treatment, ECOG performance status, CA at diagnosis, extramedullary disease, new bone lesions (X-ray), refractory status and severe toxicities during 1L treatment (any grade 3 or 4 toxicity). All predictor values used in the risk stratification algorithm were measured at initiation of 2L treatment with the exception of CA, LDH and Sβ2M; the latter two predictors were measured at both diagnosis and initiation of 2L treatment. A corresponding frailty score (defined by age and ECOG performance status) and aggressiveness score (defined by all the identified predictors specifically linked to the disease characteristics) were calculated for each patient.

The data used to develop the risk stratification algorithm were sourced from the Czech RMG.[18]

Data analysis

Data for the validation were derived from patient medical chart audits in France, Germany and the UK. A bespoke retrospective chart review cohort study was designed to collect all of the real-world data required to validate the risk stratification algorithm. Participating physicians were oncologists, oncohaematologists and haematologists; in total, 60 physicians participated from France, 70 from Germany and 50 from the UK. Patients with symptomatic

MM were documented if they were initiated on 2L anti-tumour drug treatment during 2013 (providing sufficient follow-up to collect survival outcomes). Relevant data were abstracted onto a study-specific case report form during the second and third quarters of 2017. Data from the individual countries were examined descriptively for population differences and pooled for the purpose of validation.

The baseline period was defined as the time between diagnosis and the initiation of 2L treatment. Patients were followed from diagnosis to death, end date of study inclusion (if not deceased) or date of last contact (if lost to follow-up). Recorded outcomes included OS, progression-free survival, time to disease progression and treatment response.

Real-world data were analysed on a descriptive basis. Continuous variables were summarised using descriptive statistics (number, mean, SD, median, minimum and maximum values). Categorical variables were reported as frequency counts and the percentage of individuals in corresponding categories. Survival outcomes were summarised with Kaplan-Meier curves and in terms of median (95% CI) survival and the restricted mean survival.

Multiple imputation was performed for missing values,[19] but only for predictors in the risk stratification algorithm, and with the exception of CA. No outcomes data were imputed. Owing to the lack of methods by which Cox model performance measures may be pooled, five rounds of imputation were conducted and data in the third imputed set were analysed.

Validation procedure

The performance of the risk stratification algorithm was evaluated in terms of the predictive performance of the Cox regression model and stratification of patients for OS.

A detailed description of the statistical analysis on the performance methods has been described previously.[15] The performance of the Cox regression model for predicting the risk of death was assessed according to the extent to which the variance in OS was explained by the selected predictors (Nagelkerke's R^2), as well as the discriminative power (Harrell's concordance index; C-index): point estimate (95% CI).[20 21] The discriminative power of the Cox regression model was regarded as accurate if the C-index was ≥0.70.

The performance of the risk stratification algorithm for stratifying patients in groups by OS was analysed by evaluating OS by risk group (using Kaplan-Meier curves) and HRs comparing risk groups. A comparison of the HRs across validation datasets was used to indicate the extent to which scores could be reliably interpreted in different countries/settings.

Patient frailty (based on age and ECOG performance status) and disease aggressiveness (based on all other predictors in the model) in different risk groups and the relationship between them across risk groups were also investigated.

Patient and public involvement statement

Data for the validation cohort were derived retrospectively from patient medical chart audits. As such, patients and the general public were not involved in this study.

RESULTS

Retrospective chart review

Chart data from a total of 998 patients were collected (France, 386; Germany, 344; UK, 268). The characteristics of these patients at diagnosis, including the 16 parameters used in the risk stratification algorithm, are summarised by country in table 1. Additional parameters not included in the risk stratification algorithm are described in online supplementary table S1. Certain between-country differences were observed (not compared statistically), such as a tendency for lower ISS and higher ECOG performance status scores in France versus Germany or the UK. In the validation cohort, 46.0% of patients had a prior stem cell transplant (SCT) and 11.5% of patients had an SCT at both 1L and 2L. A summary of the characteristics that were included in the risk stratification algorithm (pooled across the three countries) has been previously published alongside data from the original RMG dataset, and a number of discrepancies highlighted.[16] Proportionally more patients in the validation cohort than in the Czech development cohort had elevated LDH levels, hypercalcaemia and higher bone-marrow plasma cell counts and bone lesions, all of which are associated with an increased overall risk of death. However, this was mitigated to some extent by the fact that proportionally fewer patients in the validation cohort than in the Czech development cohort had proven refractory to thalidomide or had experienced grade 3–4 toxicity during 1L treatment.[16]

Cox model performance analysis

The point estimate for the C-index in the validation cohort was 0.715 (95% CI 0.690 to 0.734—a score of 0.5 represents total random predictions; a score of 1 represents a perfectly discriminating model; a good discriminating model has a score of >0.70). The R^2 value for the validation set was 0.283 (possible scores range between 0 and 1 for a model that explains 0%–100% of the observed variation) based on 437 events observed in 998 patients. For comparison, the R^2 for the Czech development cohort was 0.253 (737 events in 1418 patients).

Stratification of patients

The distributions of patients across the four risk groups in the validation cohort and the original Czech RMG dataset have been described previously[16]; distribution was a little more even in the validation cohort than the RMG dataset, which was more skewed towards the lower risk groups. Patient characteristics at diagnosis and at the initiation of 2L treatment by risk group are summarised

Table 1 Patient characteristics included in the risk stratification algorithm by country

Characteristic	Total (n=998)	France (n=386)	Germany (n=344)	UK (n=268)	HR*	P value†	
Frailty score							
Parameters included in risk stratification algorithm As measured at initiation of 2L treatment unless otherwise stated					1.015		Continuous variable
Age (years)						0.0002	
≤65	456 (45.7)	159 (41.2)	143 (41.5)	154 (57.5)			
66–75	369 (37.00)	145 (37.6)	146 (42.4)	78 (29.1)			
>75	173 (17.3)	82 (21.3)	55 (16.0)	36 (13.5)			
Mean (SD)	66.4 (10.2)	67.6 (10.0)	66.5 (10.1)	64.4 (10.1)			
Median (IQR)	67 (59–74)	68 (60–75)	68 (59–74)	64 (57–73)			
ECOG performance status							Categorical variable
0	144 (14.4)	44 (11.4)	47 (13.7)	53 (19.8)	–		
1	566 (56.7)	210 (54.4)	191 (55.5)	165 (61.6)	1.667	0.0011	
2	258 (25.9)	114 (29.5)	98 (28.5)	46 (17.2)	2.123	<0.0001	
3–4	30 (3.0)	18 (4.7)	8 (2.3)	4 (1.5)	3.708	<0.0001	
Albumin (g/dL)					0.846		Continuous variable
<3.5	509 (51.0)	157 (40.7)	170 (49.4)	182 (67.9)		0.0095	
≥3.5	489 (49.0)	229 (59.3)	174 (50.6)	86 (32.1)			
Mean (SD)	3.5 (0.9)	3.7 (1.1)	3.5 (1.0)	3.3 (0.5)			
Median (IQR)	3.4 (3.0–3.8)	3.6 (3.1–3.9)	3.5 (3.0–3.9)	3.2 (2.9–3.6)			
Bone marrow plasma cell count (%)					1.008		
<20	240 (24.0)	131 (33.9)	51 (14.8)	58 (21.6)		<0.0001	
20–70	668 (66.9)	239 (61.9)	261 (75.9)	168 (62.7)			
>70	90 (9.0)	16 (4.1)	32 (9.3)	42 (15.7)			
Mean (SD)	36.6 (22.7)	16 (4.1)	32 (9.3)	42 (15.7)			
β2 microglobulin (mg/L)					1.063 (up to 5.5 mg/L)		Continuous variable with threshold
<3.5	339 (34.0)	160 (41.5)	109 (31.7)	70 (26.1)		0.0787	
3.5–5.5	405 (40.6)	165 (42.7)	122 (35.2)	118 (44.4)			
>5.5	254 (25.5)	61 (15.8)	114 (33.1)	79 (29.5)			
Mean (SD)	4.5 (2.5)	4.0 (2.1)	5.0 (3.1)	4.7 (1.9)			
Median (IQR)	3.9 (3.0–5.6)	3.7 (2.7–4.8)	4.1 (3.0–6.0)	4.2 (3.4–5.8)			
β2 microglobulin—diagnosis (mg/L)					1.090 (up to 5.5 mg/L)		
<3.5	369 (37.0)	186 (48.2)	107 (31.1)	76 (28.4)		0.0084	
3.5–5.5	389 (39.0)	138 (35.8)	140 (40.7)	111 (41.4)			
>5.5	240 (24.0)	62 (16.1)	97 (28.2)	81 (30.2)			
Mean (SD)	4.4 (2.3)	4.0 (2.3)	4.7 (2.4)	4.7 (1.9)			
Median (IQR)	3.9 (3.0–5.5)	3.6 (2.7–4.6)	4.0 (3.2–5.8)	4.4 (3.2–5.8)			
Thrombocyte count (10^9 cells/L)					0.995 (up to 150×10^9 cells)		
>100	867 (86.9)	343 (88.9)	271 (78.8)	253 (94.4)		<0.0001	
≤100	131 (13.1)	43 (11.1)	73 (21.2)	15 (5.6)			
Mean (SD)	180.3 (79.4)	176.8 (70.4)	168.4 (90.1)	200.7 (72.5)			
Median (IQR)	176 (121–212)	177.0 (125.0–209.0)	154.0 (110.0–207.0)	195 (146.0–245.0)			

Continued

Table 1 Continued

Characteristic	Total (n=998)	France (n=386)	Germany (n=344)	UK (n=268)	HR*	P value†	
LDH (U/L)					2.080 (>ULN)		Continuous variable with clinically established cut-off
Below ULN‡	817 (81.9)	320 (82.9)	302 (87.8)	195 (72.8)			
Above ULN‡	180 (18.1)	66 (17.1)	42 (12.2)	73 (27.2)		<0.0001	
Mean (SD)	303.5 (146.9)	288.8 (137.4)	274.7 (100.3)	361.7 (188.3)			
Median (IQR)	264.0 (201.0–362.0)	233.0 (200.0–361.0)	247.0 (200.0–320.0)	314 (219.0–415.0)			
LDH at diagnosis (U/L)					1.297 (>360 U/L)		
Below ULN‡	832 (83.4)	324 (83.9)	304 (88.4)	218 (76.1)			
Above ULN‡	166 (16.6)	62 (16.1)	40 (11.6)	64 (23.9)		0.0904	
Mean (SD)	296.0 (157.7)	282.2 (147.5)	275.3 (132.0)	342.4 (189.5)			
Median (IQR)	246.0 (200.0–350.0)	230.0 (195.0–345.0)	240.0 (200.0–328.0)	297.0 (207.0–397.0)			
Hypercalcaemia§					1.406 (>2.75 mmol/L)		
No	799 (80.1)	316 (82.1)	246 (71.5)	252 (88.1)			
Yes	199 (19.9)	69 (17.9)	98 (28.5)	32 (11.9)		0.0422	
Mean (SD)	3.4 (2.6)	3.3 (2.6)	4.0 (3.0)	2.9 (1.7)			
Median (IQR)	2.4 (2.2–2.9)	2.4 (2.3–2.7)	2.4 (2.2–3.9)	2.5 (2.3–2.8)			
Time to initiation of 2L treatment (months)					1.112 (≤24 months)		
>24	467 (46.8)	224 (58.0)	133 (38.7)	110 (41.0)			
≤24	531 (53.2)	162 (42.0)	211 (61.3)	158 (59.0)		0.2858	
CAs at diagnosis							Categorical variable
Standard risk	332 (33.3)	122 (31.6)	143 (41.6)	67 (25.0)	–		
High risk	209 (20.9)	74 (19.2)	68 (19.8)	67 (25.0)	1.643	0.0067	
NA¶	457 (45.8)	190 (49.2)	133 (38.7)	134 (50.0)	1.081	0.6299	
Extramedullary disease							
No	879 (88.1)	326 (84.5)	311 (90.4)	242 (90.4)	–		
Yes	119 (11.9)	60 (15.5)	33 (9.6)	26 (9.7)	2.331	<0.0001	
Refractory to previous treatment							
Non-refractory/ refractory to other regimens without new drugs	787 (78.9)	309 (80.1)	270 (78.5)	208 (77.6)	–		
Refractory to bortezomib	123 (12.3)	53 (13.7)	47 (13.7)	23 (8.6)	1.533	0.0006	
Refractory to thalidomide	1 (0.1)	–	–	1 (0.4)	1.186	0.1446	
Refractory to other regimens with new drugs**,††	87 (8.7)	24 (6.2)	27 (7.8)	36 (13.4)	1.427	0.0776	
New bone lesions							
No new lesions	396 (39.7)	206 (53.4)	107 (31.1)	83 (31.0)	–		
>2 lesions at diagnosis and at 2L, or new lesions	602 (60.3)	180 (46.6)	237 (68.9)	185 (69.0)	1.271	0.0049	
Severe toxicities during/before 1L treatment (highest grade experienced)							
0–2	879 (88.1)	356 (92.7)	281 (81.7)	240 (89.6)	–		
3–4	119 (11.9)	28 (7.3)	63 (18.3)	28 (10.4)	1.145	0.0797	

Data are n (%) unless otherwise stated.
*HRs for calculating individual risk score (based on Cox regression analysis of development cohort).
† The p values are associated with the effects of the predictors in the risk stratification algorithm (based on Cox regression analysis of the development cohort) and not related to the comparison of countries.
‡Upper limit of normal (ULN) in the Czech data was estimated to be 360 U/L.
§Hypercalcaemia in the Czech data was defined as >2.75 mmol/L.
¶Missing values were not imputed for cytogenetic abnormalities (CAs).
**'New' drugs include carfilzomib, daratumumab, elotuzumab, ixazomib, panobinostat, pomalidomide and thalidomide.
††Refractory to other regimens with new drugs—includes bortezomib plus thalidomide, lenalidomide only, bortezomib plus lenalidomide and lenalidomide plus thalidomide.
ECOG, Eastern Cooperative Oncology Group; 1L, first line; 2L, second line; LDH, lactate dehydrogenase; NA, not available.

Table 2 Base case analysis (imputed dataset 3) of overall survival in the validation cohort and in the development cohort

	Validation cohort			Development cohort		
Risk group	n	Median overall survival (months; 95% CI)	HR (95% CI)	n	Median overall survival (months; 95% CI)	HR (95% CI)
1	178	NA (NA–NA)	Reference	351	61.6 (51.7 to 71.4)	Reference
2	345	NA (NA–NA)	1.868 (1.234 to 2.827)	596	29.6 (26.6 to 32.6)	2.24 (1.78 to 2.83)
3	249	39.8 (32.7 to 46.7)	4.613 (3.091 to 6.883)	318	14.2 (11.3 to 17.1)	4.30 (3.38 to 5.49)
4	226	16.2 (13.7 to 21.0)	8.514 (5.735 to 12.64)	153	5.9 (4.4 to 7.5)	10.88 (8.32 to 14.23)
Pooled	998	NA	NA	1418	27.6 (25.0 to 30.0)	NA

NA, not available.

in online supplementary table S2. Differences between risk groups were identified for parameters such as age (a trend for increasing mean age with risk group) and transplant status (fewer transplants for patients in groups 3 or 4). Both ISS at diagnosis and ECOG performance status scores at 2L initiation showed a tendency for higher values with increasing risk group.

OS by risk group in the validation dataset is shown in table 2 (alongside data for the Czech development cohort for comparison). It is notable that OS was considerably longer in the validation dataset than in the original Czech development cohort. As was the case with the original Czech development cohort, there was clear differentiation between HRs for OS between the risk groups in the validation dataset. The HRs for differences in OS between patients in group 1 and in groups 2, 3 and 4 were 1.87, 4.61 and 8.51, respectively.

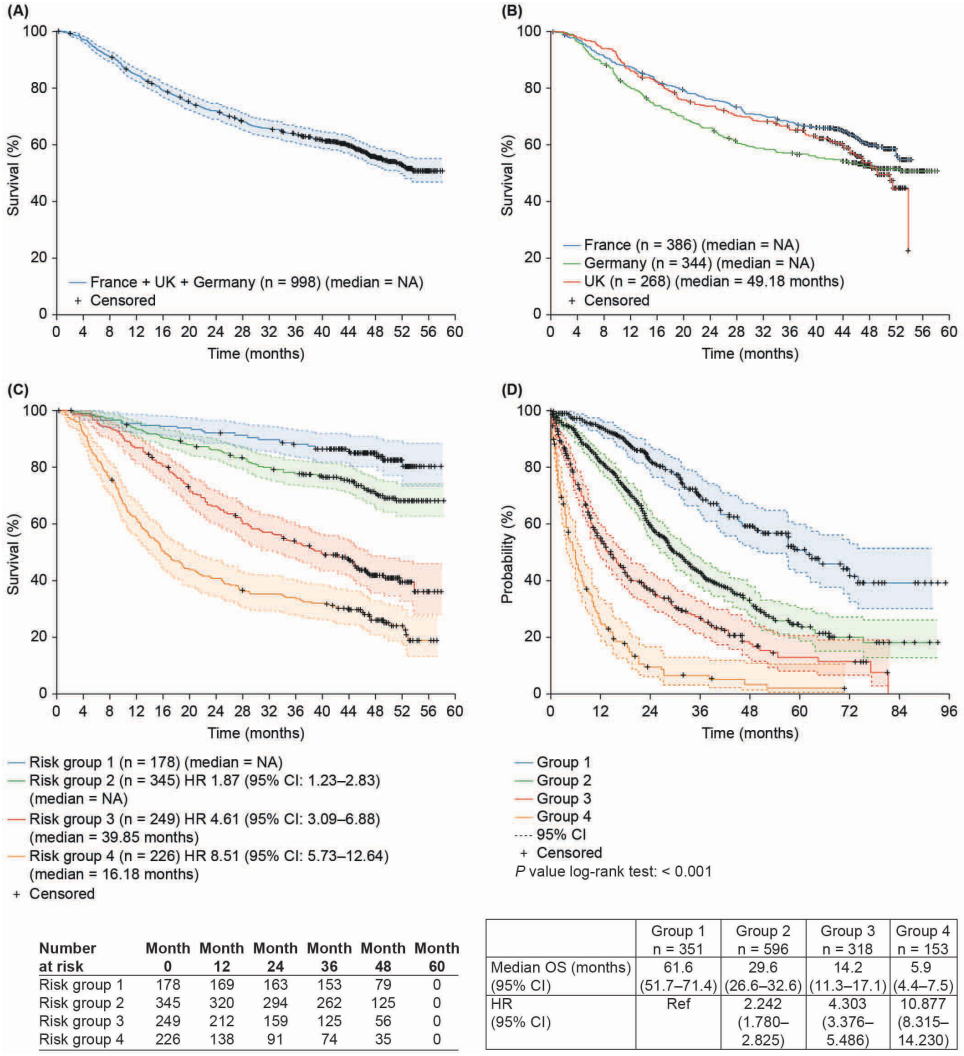
Kaplan-Meier plots for OS for the pooled validation set (figure 1A) and by country (figure 1B), as well as by risk group for the validation (figure 1C) and development (figure 1D) cohorts, are shown. The OS curve for the pooled set is immature, as the median was not reached during follow-up (figure 1A). This was also the case in the French and German cohorts (figure 1B) and, as noted above, in risk groups 1 and 2 (figure 1C). Median OS in the UK cohort was 49.2 months.

Disease aggressiveness and patient frailty were key components of classification in the original development of the novel risk stratification algorithm (figure 2). A scatterplot of disease aggressiveness scores versus frailty scores in the overall validation cohort and Czech development cohort is shown in figure 2A and B, respectively, demonstrating clear stratification by risk group in each cohort. Considering the groups in sequence from 1 to 4 (online supplementary figure S1A–C), it seems that frailty scores were spread quite broadly from group 2 onwards (online supplementary figure S1B), whereas disease aggressiveness scores were only markedly spread in group 4 (online supplementary figure S1C). Notably, consideration of frailty and disease aggressiveness in group 4 patients shows that the frailest patients are not necessarily experiencing the most aggressive disease, and vice versa. It could be expected that patients with both a high frailty and high aggressiveness score may not have been able to survive to reach 2L therapy.

Univariate Cox models showed that the HRs of death associated with each unit increase in the total risk scores, aggressiveness scores and frailty scores for the validation cohort were 1.018, 1.101 and 1.341, respectively (table 3). Some differences were observed between the frailty and aggressiveness scores among countries; for example, frailty scores were higher in the UK cohort than in the French or German cohort (1.674, 1.302 and 1.428 for UK, France and Germany, respectively).

DISCUSSION

The novel risk stratification algorithm was developed using data from one of the largest existing registries of patients with MM and monoclonal gammopathies of unknown significance. The RMG contains detailed information on a large number of patient characteristics and disease-related parameters recorded at diagnosis and at first relapse. It has mature OS data and is representative of the national and international patient populations. In order to ensure generalisability of this tool to all patients with relapsed MM, however, it was important to validate the tool in an independent cohort. We therefore conducted a bespoke retrospective chart review cohort study in order to collate data from patients in France, Germany and the UK to collect data for this purpose. The pooled validation population from these three countries was around two-thirds of the size of the development cohort and demonstrated significant differences in disease characteristics.[16] Despite these differences, results were consistent between the two datasets, which highlights the ability of



Number at risk	Month 0	Month 12	Month 24	Month 36	Month 48	Month 60
Risk group 1	178	169	163	153	79	0
Risk group 2	345	320	294	262	125	0
Risk group 3	249	212	159	125	56	0
Risk group 4	226	138	91	74	35	0

	Group 1 n = 351	Group 2 n = 596	Group 3 n = 318	Group 4 n = 153
Median OS (months) (95% CI)	61.6 (51.7–71.4)	29.6 (26.6–32.6)	14.2 (11.3–17.1)	5.9 (4.4–7.5)
HR (95% CI)	Ref	2.242 (1.780–2.825)	4.303 (3.376–5.486)	10.877 (8.315–14.230)

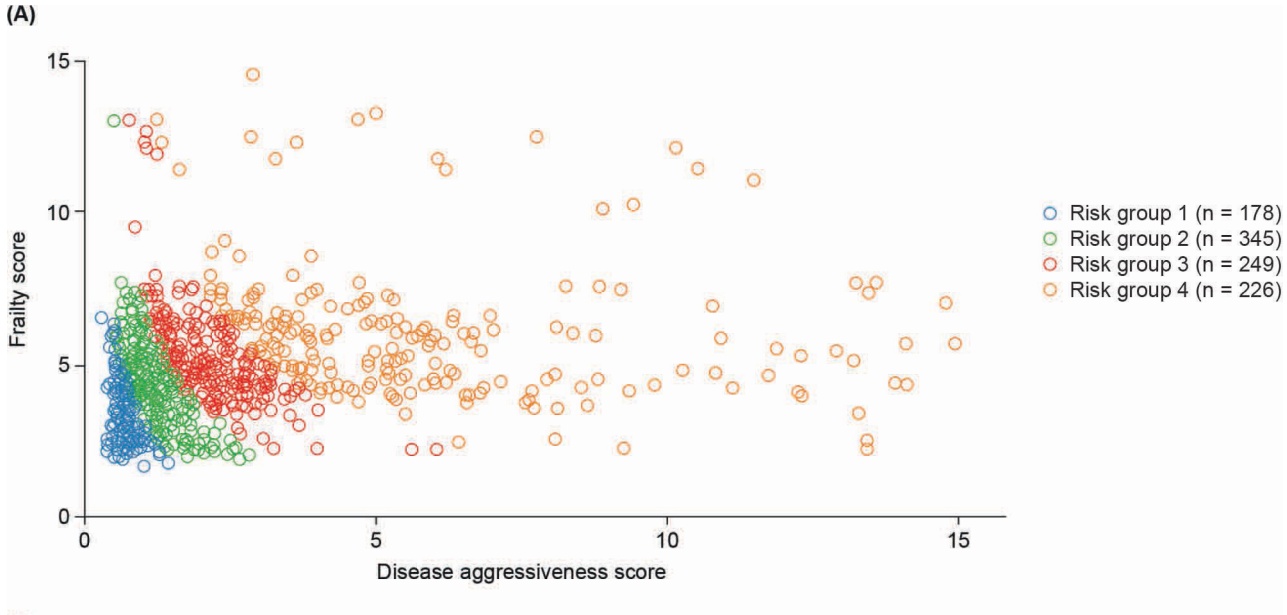
Figure 1 Kaplan-Meier plots for overall survival in (A) the pooled validation cohort, (B) by country and (C) by risk group (imputed dataset 3); (D) by risk group in the development cohort. NA, not available; OS, overall survival; Ref, reference value.

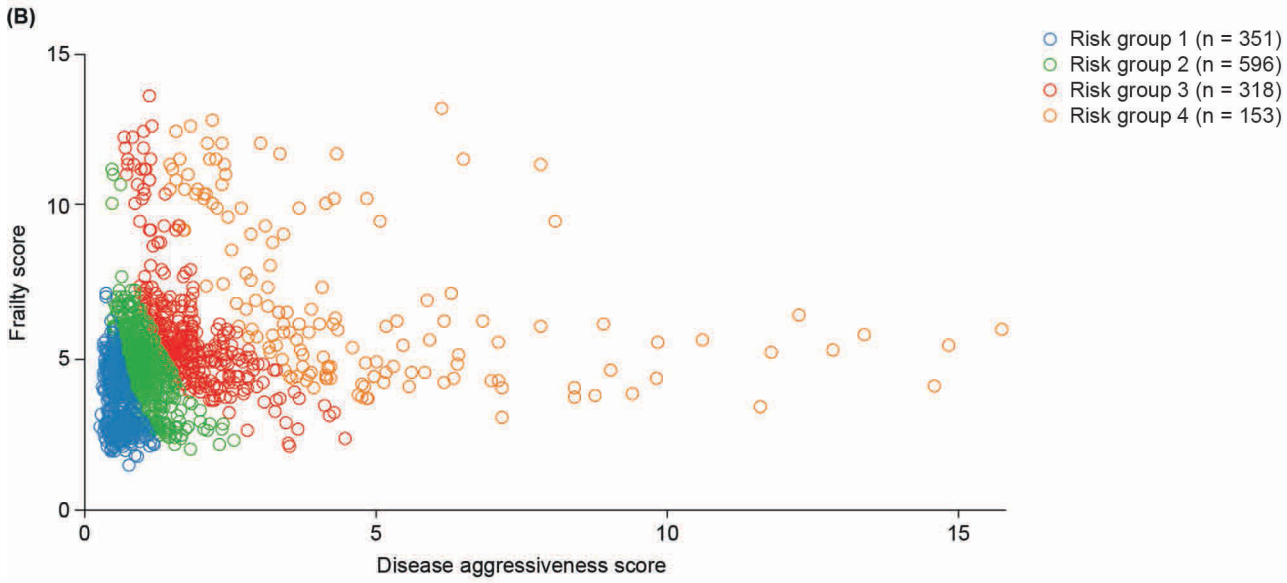
the risk stratification algorithm to perform independent of patient heterogeneity.

Differences were identified between the patient characteristics at diagnosis across the three countries in the validation dataset. ISS suggested, for example, that patients in France had less severe disease than in Germany or the UK, while patients in the UK and France were more likely to have undergone transplantation than those in Germany. The heterogeneity of the validation population overall supports its value in demonstrating the generalisability of the risk stratification algorithm for use in practice. It is interesting to note that there was not complete agreement between the scoring systems; nearly one-fifth of patients classified as ISS group III fell into risk groups 1 and 2, for example, while 13% of those in the lowest risk category based on ISS at diagnosis were identified by the risk stratification algorithm as being in the top two risk groups.

The C-index for the validation dataset (0.715), which was similar to the Czech development cohort (0.723),[15] indicated that the Czech development cohort demonstrated accurate discriminative power across the validation dataset; the discriminative power of the Cox regression model could be regarded as accurate if the C-index was ≥0.70.[21] The Nagelkerke's R^2 values were similar, suggesting that the Czech development model was able to explain variance in OS to a comparable extent in the two cohorts.

With regards to stratification of the risk groups, the Kaplan-Meier curves were shown to separate from early on in both the development and validation cohorts. There was little overlap of the Kaplan-Meier estimates for OS in the four groups during follow-up (figure 1C), or of the 95% CIs for HRs comparing each risk group versus group 1, indicating how well the four groups were differentiated. A similar trend was observed in the development cohort (figure 1D). As risk groups increased, the HRs doubled in both the development and validation cohorts, demonstrating consistency in risk stratification of patients. In the Czech development cohort, median OS reduced by half as risk groups increased; in the validation cohort, median OS was not reached in groups 1 and 2 and





Validation RSA	Mean	Median (IQR)	Range
Total risk score	13.3	6.6 (3.6–14.1)	0.8–232.7
Frailty score	4.8	4.7 (3.9–5.6)	1.7–14.5
Aggressiveness score	2.5	1.4 (0.9–2.7)	0.3–31.6

Czech RSA	Mean	Median (IQR)	Range
Total risk score	7.9	4.8 (3.1–8.7)	0.8–94.1
Frailty score	5.0	4.8 (4.0–5.5)	1.5–13.7
Aggressiveness score	1.5	1.0 (0.7–1.6)	0.3–18.7

Figure 2 Frailty versus disease aggressiveness score by risk group in the pooled dataset: (A) validation cohort; (B) development cohort. RSA, risk stratification algorithm.

reduced by half in the higher risk groups. Adjustment for 2L treatment in the validation dataset had little impact on HRs for OS (data not shown). To put the HR results into context, it is worth considering corresponding data from the R-ISS. Palumbo *et al* reported HRs of 3.68 and 9.95 for R-ISS groups II and III, respectively, versus group I.[9]

Comparing the same risk groups in the validation and Czech RMG populations revealed notable differences in median OS in the two cohorts. However, detailed consideration of outcomes in the two populations was impeded, to an extent, by the fact that median OS was not reached in several groups in the validation population during the

Table 3 Univariate Cox model score in the pooled validation cohort and by country

	Total risk score	Aggressiveness score	Frailty score
Pooled	1.018 (1.015 to 1.020)	1.101 (1.083 to 1.120)	1.341 (1.287 to 1.398)
France	1.023 (1.017 to 1.029)	1.149 (1.101 to 1.198)	1.302 (1.224 to 1.384)
Germany	1.020 (1.015 to 1.024)	1.106 (1.077 to 1.136)	1.428 (1.314 to 1.552)
UK	1.017 (1.012 to 1.021)	1.090 (1.061 to 1.120)	1.674 (1.491 to 1.880)

Data are effect of score in terms of HRs (95% CI).

60-month follow-up. The higher OS values in the validation cohort may imply a somewhat healthier patient population than the Czech RMG cohort, but in terms of deaths related to relapsed MM that conclusion is not supported by some of the differences in predictive characteristics already described. Variation in OS would be expected in different populations, owing to modifications in treatment regimens or in other factors such as comorbidities and lifestyle, but importantly the trend in survival expectations associated with the risk stratification algorithm risk-group stratification was common to both populations, as can be seen by the similarity of the HRs in each risk group when estimated relative to group 1 patients in the same population (table 3). Median OS in all risk groups would be expected to increase as treatment regimens improve, but in the absence of a cure, the stratification process will be clinically useful when paired with an understanding of outcomes associated with different treatments in each risk group. This validation process can boast a number of design strengths. Critically, the use of a large, heterogeneous validation population, with demonstrated variability in populations across three different countries, ensured robustness of the validation approach. This population differed from the Czech development population with which the tool was initially developed, so the validation represented a good test of generalisability. The chart data were also recent, in order to maximise the relevance of the validation to current clinical practice.

There were also, inevitably, certain limitations of the study. In terms of the representativeness of the sample, as mentioned earlier it appears that the external validation cohort may have been healthier than the original Czech RMG population, based on OS estimates. It is intended that the risk stratification algorithm will be validated in further groups of patients with relapsed MM in order to ensure that the tool's performance has been evaluated across a wide spectrum of patients. Validation in Greek[22] and Slovakian populations is underway; these feature relatively short OS measurements, and thus will help to address the issue of the disproportionately healthy current validation cohort. From a methodological perspective, the current analysis used a mix of Cox model performance (R^2, C-index) and simply computed HRs comparing risk groups, as there are no established measures for assessing risk stratification algorithm performance. Finally, in terms of the data obtained, there were a high number of missing values for certain variables, such as Sβ2M at diagnosis and LDH and albumin at 2L. These particular parameters are not typically used to guide treatment, and thus are rarely tested in routine practice; such issues are difficult to avoid in studies relying on real-world data.

CONCLUSION

The risk stratification algorithm has now been validated in an independent European cohort. Consistent results in risk stratification have been demonstrated between the validation and development cohorts, in terms of HR and median OS differences across risk groups. This is the first specifically designed patient risk assessment tool that combines both frailty and aggressiveness metrics into a single score. Patient-specific risk as assessed by this tool can be used as a prognostic factor to tailor management strategies for patients, based on burden of disease, capacity to benefit, urgency to treat and to aid decisions on the intensity of therapy.

Author affiliations
[1]Department of Haematooncology, University Hospital Ostrava, Ostrava, Czech Republic
[2]Amgen Europe GmbH, Rotkreuz, Switzerland
[3]Department of Haematology, University of Leuven, Leuven, Belgium
[4]Amgen, Uxbridge, UK
[5]Department of Internal Medicine, University Hospital Heidelberg, Heidelberg, Germany
[6]Pharmerit International, Rotterdam, The Netherlands
[7]Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK

Acknowledgements Medical writing support was provided by Kim Allcott PhD of Oxford PharmaGenesis, Oxford, UK, and was funded by Amgen Europe GmbH

Contributors SG-M, LD and WB designed and performed the research study. All authors analysed the data and contributed to writing the paper by providing guidance and comments on its content and they critically revised the manuscript and agreed to the final version.

Funding This work was funded by Amgen Europe GmbH

Competing interests RH has received research funding from Amgen and Celgene, consultancy fees from Amgen, Celgene and Takeda, and honoraria from Amgen, Bristol-Myers Squibb and Janssen. SG-M, ZS and MC are employees of Amgen Europe and stockholders in Amgen. MD has received research funding from Celgene and Janssen, and honoraria from Amgen, Bristol-Myers Squibb, Celgene, Janssen and Takeda. LD is an employee of Amgen and a stockholder in Amgen. MSR has received research funding from Amgen and Novartis, consultancy fees from Amgen, Bristol-Myers Squibb, Celgene, Takeda and Novartis, and has participated in advisory boards for Celgene, Bristol-Myers Squibb, Amgen and Janssen. WB is an employee of Pharmerit International, which received funding

from Amgen to conduct this research. AB has received consultancy fees from Amgen in relation to the work reported here.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval The retrospective chart review used in this study was approved by the Ethics commission of the Medical Faculty Heidelberg, Germany.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

ORCID iD

Roman Hájek http://orcid.org/0000-0001-6955-6267

REFERENCES

1 Dimopoulos MA, Terpos E, Niesvizky R, *et al*. Clinical characteristics of patients with relapsed multiple myeloma. *Cancer Treat Rev* 2015;41:827–35.
2 Moreau P, de Wit E. Recent progress in relapsed multiple myeloma therapy: implications for treatment decisions. *Br J Haematol* 2017;179:198–218.
3 Yong K, Delforge M, Driessen C, *et al*. Multiple myeloma: patient outcomes in real-world practice. *Br J Haematol* 2016;175:252–64.
4 Moreau P. The future of therapy for relapsed/refractory multiple myeloma: emerging agents and novel treatment strategies. *Semin Hematol* 2012;49 Suppl 1:S33–46.
5 Raab MS, Cavo M, Delforge M, *et al*. Multiple myeloma: practice patterns across Europe. *Br J Haematol* 2016;175:66–76.
6 Kim SM, Kim MJ, Jung HA, *et al*. Comparison of the Freiburg and Charlson comorbidity indices in predicting overall survival in elderly patients with newly diagnosed multiple myeloma. *Biomed Res Int* 2014;2014:1–11.
7 Greipp PR, San Miguel J, Durie BGM, *et al*. International staging system for multiple myeloma. *J Clin Oncol* 2005;23:3412–20.
8 Hari PN, Zhang M-J, Roy V, *et al*. Is the International staging system superior to the Durie-Salmon staging system? A comparison in multiple myeloma patients undergoing autologous transplant. *Leukemia* 2009;23:1528–34.
9 Palumbo A, Avet-Loiseau H, Oliva S, *et al*. Revised international staging system applied to real world multiple myeloma patients. *J Clin Oncol* 2015;33:2863–9.
10 Jimenez-Zepeda VH, Duggan P, Neri P, *et al*. Revised international staging system applied to real world multiple myeloma patients. *Clin Lymphoma Myeloma Leuk* 2016;16:511–8.
11 Bustoros M, Mouhieddine TH, Detappe A, *et al*. Established and novel prognostic biomarkers in multiple myeloma. *Am Soc Clin Oncol Educ Book* 2017;37:548–60.
12 Tandon N, Rajkumar SV, LaPlant B, *et al*. Clinical utility of the revised international staging system in unselected patients with newly diagnosed and relapsed multiple myeloma. *Blood Cancer J* 2017;7:e528.
13 Hájek R, Jarkovsky J, Bouwmeester W, *et al*. The value of risk stratification tools in multiple myeloma (Mm) in the real-world: validation of the Revised- international staging system (R-ISS) at initiation of treatment and relevance of the ISS and the R-ISS for risk stratification in the relapsed setting using data from the Czech registry of monoclonal gammopathies (RMG). *Blood* 2016;128:2418.
14 Hájek R, Jiri J, Walter B, *et al*. Predictors of overall survival (OS) in patients with multiple myeloma (Mm) initiating first- and second-line treatment in the Czech Republic. *Blood* 2016;128:3607.
15 Bouwmeester W, Briggs A, van Hout B, *et al*. Methodology of a novel risk stratification algorithm for patients with multiple myeloma in the relapsed setting. *Oncol Ther* 2019;7:141–57.
16 Hájek R, Delforge M, Raab MS, *et al*. Development and validation of a novel risk stratification algorithm for relapsed multiple myeloma. *Br J Haematol* 2019;187:447–58.
17 Gonzalez-McQuire S, Dimopoulos M-A, Weisel K, *et al*. Development of an initial conceptual model of multiple myeloma to support clinical and health economics decision making. *MDM Policy Pract* 2019;4:238146831881425.
18 Radocha J, Pour L, Spicka I, *et al*. Registry of monoclonal gammopathies (RMG) in the Czech Republic. *Blood* 2015;126:4514.
19 van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999;18:681–94
20 Taylor JMG, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. *Clin Cancer Res* 2008;14:5977–83.
21 Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172:971–80.
22 Terpos ED MA, Kastritis E, Roussou M. *External validation of the multiple myeloma risk-stratification algorithm in a real-world Greek data set.Poster presented at the European hematology*. Association 23rd annual meeting 14-17 June, 2018.